Assessment

Student *assessment* involves collection of information about their learning

Evaluation refers to a judgement. During grading, a judgement is made on the assessed student.

Purpose of assessment

- Feedback on *instruction* to the teacher
- Feedback on *learning* to the student

Formative assessment

- *Grading* of the student

Summative asssessment

Some Methods of Assessing Students

Information on student understanding and performance can be gathered in many different ways:

- a) Through questions asked by the teacher during a plenary session (teacher lectures/discusses but also asks questions).
- b) By looking at homework or seatwork assignments (teacher gives students seatwork and uses the time to go around to check homework or see how the seatwork is going).
- c) By interviewing some students during the lesson while other students have been given seatwork or small group work. The purpose of this interviewing is to go more in depth and probe students' conceptual understanding.

Some Methods of Assessing Students (contd)

- d) By having students prepare/make a presentation.
- e) Through other student products (posters, role plays, reports, homework, problem solving).
- f) Through short or long tests (essay, multiple choice, other formats).
- g) Observing students while they carry out assignments, for example, observing lab work , role-plays or other student activity.

Some Methods of Assessing Students (contd)

- **Please note that tests are only one of many ways to find out about student progress, but also note that *not* all the methods above are suitable for grading purposes.**

Methods of learning (contd)

- Feedback on *learning* to the student: Students can sit through many lessons and not realize that they do not master the objectives. They may not even realize *what* they are supposed to learn. They may feel confident and not be aware of misconceptions they hold. Sometimes it is only during long exams students become aware that their understanding is insufficient or what it is that the teacher wants them to learn. It is important that students find out very quickly whether or not they understand and what it is that they do not understand and what it is that they should understand. In other words, assessment can *guide* the learning process, it can communicate to them what they are supposed to learn and what they understand and not understand.

***Formative* assessment**

- is the collection of performance data from students in order to improve instruction and learning. Formative assessment is done while *teaching and learning are in process*, at a time when information from students can still be used to improve the learning and instruction. Formative assessment should be used to *guide* and *improve* the learning process.
- One way to assess student performance is to use a *diagnostic* test. The purpose is to assess student understanding without giving a grade.

***Summative* assessment**

- *assessment* is done at the end of the instruction and learning process, for example in a long exam at the end of a chapter or grading period. Then it is too late to still influence the learning and instruction for that topic…it is the end. The summative assessment is used mainly for grading students. Tests used in summative assessment are typically *achievement* tests and they are graded.

Measuring Pupil Achievement

- In today's policy environment, testing has become a critical component of education reform. Policy makers and education administrators often view test scores as a measure of educational quality and use test scores to hold schools accountable for teacher performance. ***Continuous assessment***, an alternative or supplement to high stakes testing of pupil achievement, offers a methodology for measuring pupil performance and using those findings to improve the success of pupils.

Continuous assessment

- is a classroom strategy implemented by teachers to ascertain the knowledge, understanding, and skills attained by pupils.
- Teachers administer assessments in a variety of ways over time to allow them to observe multiple tasks and to collect information about what pupils know, understand, and can do. These assessments are curriculum-based tasks previously taught in class.
- ***Continuous assessment*** occurs frequently during the school year and is part of regular teacher-pupil interactions.

Continuous assessment (contd)

- Pupils receive feedback from teachers based on their performance that allows them to focus on topics they have not yet mastered.
- Teachers learn which students need *review and remediation* and which pupils are ready to move on to more complex work. Thus, the results of the assessments help to ensure that all pupils make learning progress throughout the school cycle thereby increasing their academic achievement.

*Some benefits of **continuous assessment***

- The **continuous assessment** process is much more than an examination of pupil achievement. **Continuous assessment** is also a powerful diagnostic tool that enables pupils to understand the areas in which they are having difficulty and to concentrate their efforts in those areas. Some of its benefits are:

*Some benefits of **continuous assessment*** (contd)

- *allows teachers to monitor the impact of their lessons on pupil understanding*. Teachers can modify their pedagogical strategies to include the construction of remediation activities for pupils who are not working at the expected grade level and the creation of enrichment activities for pupils who are working at or above the expected grade level. Hence, the **continuous assessment** process *supports a cycle of self-evaluation and pupil-specific activities by both pupils and teachers*.
- Frequent interactions between pupils and teachers means that teachers know the strengths and weaknesses their learners. These exchanges foster a pupil-teacher relationship based on individual interactions. Pupils learn that the teacher values their achievements and that their **assessment** outcomes have an impact on the instruction that they receive. One-to-one communication between the teacher and the pupil can motivate pupils to continue attending school and to work hard to achieve higher levels of mastery.

*Some benefits of **continuous assessment*** (contd)

- In **continuous assessment**, teachers assess the curriculum as implemented in the classroom. It also allows teachers to evaluate the effectiveness of their teaching strategies relative to the curriculum, and to change those strategies as dictated by the needs of their pupils.
- In addition, **continuous** assessments provide information on achievement of particular levels of skills, understanding, and knowledge rather than achievement of certain marks or scores. Thus, **continuous assessment** enables pupils to monitor their achievement of grade level goals and to visualize their progress towards those goals before it is too late to achieve them.

Terminal and continuous assessment - two contrasting paradigms

- To discuss this (see handout, pp8-9)

Advantages of continuous assessment

- *Continuous assessment can provide much more extensive syllabus coverage than terminal assessment*; indeed, in some cases (e.g., competence-based courses) it covers virtually all aspects of the students' work, thus greatly increasing the face validity of the assessment process and permitting the use of tools appropriate to the workplace.
- Since it allows the use of a far wider range of assessment techniques than terminal assessment, continuous assessment can be used to test a correspondingly wider range of skills, including non-cognitive skills of various types. It thus makes it easier for tutors to match their assessment methods with the learning outcomes being assessed and to step assessment through different levels.

Advantages of continuous assessment (contd)

- *Continuous assessment places less emphasis on pure memory (particularly comparatively short-term memory) than terminal assessment, and correspondingly more emphasis on worthwhile learning in the deepest sense of the word*. True education has been described as 'what is left after the facts have been forgotten', and continuous assessment certainly facilitates such education.
- *continuous assessment encourages regular, systematic study and discourages last-minute cramming, thus rewarding students who work steadily and conscientiously throughout their courses*. It also reduces the domination of both teaching and learning by the requirements of the final examinations. It is like a film, rather than a single snapshot.

Advantages of continuous assessment (contd)

- By enabling on-going monitoring of student performance to take place, *continuous assessment can provide early warnings of which students are having problems with a course, thus enabling appropriate remedial help to be provided in time for it to do some good*.
- *Continuous assessment can provide early indicators of the likely performance of students, something that can be of great help to the students themselves - eg, in recognising that they have made a mistake in their choice of course and would be better transferring to another, or in helping them to make informed choices of routes and options*.

Advantages of continuous assessment (contd)

- Such assessment also *provides an on-going picture of how individual students develop and mature as they work their way through a course, something that can again be of considerable use to both students and staff*. It can also provide evidence of exactly what has been learned by a particular stage of the course, information that can prove extremely useful in cases where a student wishes to take an early exit award such as a Certificate or Diploma of Higher Education.
- Continuous assessment also *constitutes an extremely useful vehicle for on-going course monitoring and evaluation, providing course tutors with early warning of any problems or weaknesses, thus enabling them to take appropriate measures to improve matters*.

Advantages of continuous assessment (contd)

- It is generally agreed that continuous assessment *reduces the intense stress that many students experience when preparing for and sitting terminal examinations*
- Continuous assessment generally *provides a more natural assessment environment that is better matched to the situations in which students will find themselves working in later life, particularly if the assessment is of the 'open-book' variety.*

Disadvantages of continuous assessment

- Students undergoing continuous assessment may feel that they are continually under surveillance, and that every error that they make along the way can count against them. This can give rise to a different type of stress from that which students experience as a result of terminal assessment. Indeed, to quote Derek Rowntree of the Open University, "*Continuous assessment ensures that students now have ulcers as well as nervous breakdowns".*
- Unless continuous assessment is carefully planned and coordinated, there is a very real danger that students may be grossly over-assessed - particularly at certain times of the year, when several lecturers are asking simultaneously for assignments to be handed in.

Disadvantages of continuous assessment (contd)

- Attempts to broaden the scope of a course may be frustrated by students gearing their study solely to the requirements of the assessment procedures, thus putting students who carry out extension studies or 'read round' their subject at a disadvantage. By itself, continuous assessment does not prevent either 'strategic' or 'surface' learning.
- Continuous assessment can, if not properly managed, adversely affect the relationship between students and their tutors, with the latter being regarded with suspicion and (in some extreme cases) enmity and occasionally even introducing malpractice, as in imposing penalties for seeking help.

Disadvantages of continuousassessment (contd)

- Students may suffer from unequal availability of resources, something that is becoming increasingly important now that they are carrying out much of their work on personal computers or 'at a distance'.
- With continuous assessment, there is the perennial problem of enforcing uniform procedures such as completion dates and dealing with students who do not comply with these in a way that is seen to be fair without being either too draconian or too lax. Continuous assessment requires just as much planning as terminal assessment -more in many cases.

Disadvantages of continuous assessment (contd)

- Assessment schemes that are claimed to be based on 'continuous assessment' may turn out to be nothing more than a series of tests or 'mini examinations'. If so, such assessments remain 'unnatural' and fail to optimise problem-solving opportunities.
- Tutors need to have a high level of experience in assessment to enable them to make creative and effective use of continuous assessment (although the same could be said of terminal assessment!)

Is continuous assessment worthwhile?

- In view of these various possible disadvantages, is continuous assessment worthwhile on balance? In most cases, the answer is probably 'yes', since all the problem areas listed above can be overcome or obviated by careful planning and good practice. Guidance on how this can be done will be given later in the booklet.

Some of the forms that continuous assessment can take:

- Continuous assessment can be organised and implemented in a large number of different ways, and the actual assessment procedures can take many different forms. Some of the most widely used of these are outlined below.

Criterion-referenced tests (CRTs)

- *are intended to measure how well a person has learned a specific body of knowledge and skills*. Multiple-choice tests most people take to get a driver's license and on-the-road driving tests are both examples of criterion-referenced tests. As on most other CRTs, it is possible for everyone to earn a passing score if they know about driving rules and if they drive reasonably well.

Norm-referenced tests (NRTs)

- In contrast, norm-referenced tests (NRTs) *are made to compare test takers to each other*. On an NRT driving test, test-takers would be compared as to who knew most or least about driving rules or who drove better or worse. *Scores would be reported as a percentage rank with half scoring above and half below the mid-point*.

CRTs & NRTs

- In education, CRTs usually are made to determine whether a student has learned the material taught in a specific grade or course. An algebra CRT would include questions based on what was supposed to be taught in algebra classes. It would not include geometry questions or more advanced algebra than was in the curriculum. Almost all students who take the algebra test could pass if they were taught well if they studied enough and the test was well-made.

  On a standardized CRT (one taken by students in many schools), the passing or "cut-off" score is usually set by a committee of experts, while in a classroom the teacher sets the passing score. *In both cases, deciding the passing score is subjective, not objective*. Sometimes cut scores have been set in a way that maximizes the number of low income or minority students who fail the test. A small change in the cut score would not change the meaning of the test but would greatly increase minority pass rates.

CRTs and NRTs

Sometimes one kind of test is used for two purposes at the same time.

- *In addition* to ranking test takers in relation to a national sample of students, a NRT might be used to decide if students have learned the content they were taught.
- A CRT might be used to assess *mastery and* to *rank* students or schools based on their scores.

CRTs and NRTs (contd)

- NRTs are designed to sort and rank students "on the curve," not to see if they met a standard or criterion. Therefore, NRTs should not be used to assess whether students have met standards.
- In some cases, a CRT is made using technical procedures developed for NRTs, causing the CRT to sort students in ways that are inappropriate for standards-based decisions.

Conclusion

- If standardized tests are used at all,
- CRTs make more sense for schools than do NRTs.
- However, they should be based on relevant, high-quality standards and curriculum and should make the least possible use of multiple-choice and short-answer questions.
- As with all tests, CRTs and NRTs, no matter what they are called, should not control curriculum and instruction, and important decisions about students, teachers or schools should not be based solely or automatically on test scores.

NRTs &CRTs -INTENDED PURPOSES

- The major reason for using a norm-referenced test (NRT) is to classify students. NRTs are designed to highlight achievement differences between and among students to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers. School systems might want to classify students in this way so that they can be properly placed in remedial or gifted programs. These types of tests are also used to help teachers select students for different ability level, e.g., reading or mathematics instructional groups.
- With norm-referenced tests, a representative group of students is given the test prior to its availability to the public. The scores of the students who take the test after publication are then compared to those of the norm group.
- Because norming a test is such an elaborate and expensive process, the norms are typically used by test publishers for 7 years. All students who take the test during that seven year period have their scores compared to the original norm group.

NRTs &CRTs -INTENDED PURPOSES (contd)

- While norm-referenced tests ascertains the rank of students, criterion-referenced tests (CRTs) determine "...what test takers can do and what they know, not how they compare to others (Anastasi, 1988, p. 102). CRTs report how well students are doing relative to a pre-determined performance level on a specified set of educational goals or outcomes included in the school, district, or state curriculum.
- Educators or policy makers may choose to use a CRT when they wish to see how well students have learned the knowledge and skills which they are expected to have mastered. This information may be used as one piece of information to determine how well the student is learning the desired curriculum and how well the school is teaching that curriculum.

NRTs &CRTs -INTENDED PURPOSES (contd)

- Both NRTs and CRTs can be standardized.
- a standardized test as one that uses uniform procedures for administration and scoring in order to assure that the results from different people are comparable. Any kind of test--from multiple choice to essays to oral examinations--can be standardized if uniform scoring and administration are used.
- This means that the comparison of student scores is possible. Thus, it can be assumed that two students who receive the identical scores on the same standardized test demonstrate corresponding levels of performance. Most national, state and district tests are standardized so that every score can be interpreted in a uniform manner for all students and schools.

Criticism?

- NRTs have come under attack recently because they traditionally have purportedly focused on low level, basic skills.
- This emphasis is in direct contrast to the recommendations made by the latest research on teaching and learning which calls for educators to stress the acquisition of conceptual understanding as well as the application of skills.

Reliability

- Reliability is the consistency of your measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of your measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to remember that reliability is not measured, it is estimated.

Validity

- *Validity* is the strength of our conclusions, inferences or propositions. More formally, Cook and Campbell (1979) define it as the "best available approximation to the truth or falsity of a given inference, proposition or conclusion."

Estimating reliability

There are two ways that reliability is usually estimated: *test/retest and internal consistency.*

- **Test/Retest**
  Test/retest is the more conservative method to estimate reliability. Simply put, the idea behind test/retest is that you should get the same score on test 1 as you do on test 2. The three main components to this method are as follows:
- 1) implement your measurement instrument at two separate times for each subject;
  2) compute the correlation between the two separate measurements; and
  3) assume there is no change in the underlying condition (or trait you are trying to measure) between test 1 and test 2.

Estimating reliability (contd)

Internal Consistency
Internal consistency estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept.

- One common way of computing correlation values among the questions on your instruments is by using Cronbach's Alpha. In short, Cronbach's alpha splits all the questions on your instrument every possible way and computes correlation values for them all (we use a computer program for this part). In the end, your computer output generates one number for Cronbach's alpha - and just like a correlation coefficient, the closer it is to one, the higher the reliability estimate of your instrument. Cronbach's alpha is a less conservative estimate of reliability than test/retest.

Difference between *test/retest* and *internal consistency*

- The primary difference between test/retest and internal consistency estimates of reliability is that test/retest involves two administrations of the measurement instrument, whereas the internal consistency method involves only one administration of that instrument.

Types of Validity

There are *four* types of validity commonly examined in social research.

- **Conclusion validity** asks is there a relationship between the program and the observed outcome? Or, in our example, is there a connection between the attendance policy and the increased participation we saw?
- **Internal Validity** asks if there is a relationship between the program and the outcome we saw, is it a causal relationship? For example, did the attendance policy cause class participation to increase?
- **Construct validity** is the hardest to understand in my opinion. It asks if there is there a relationship between how I operationalized my concepts in this study to the actual causal relationship I'm trying to study/? Or in our example, did our treatment (attendance policy) reflect the construct of attendance, and did our measured outcome - increased class participation - reflect the construct of participation? Overall, we are trying to generalize our conceptualized treatment and outcomes to broader constructs of the same concepts.
- **External validity** refers to our ability to generalize the results of our study to other settings. In our example, could we generalize our results to other classrooms?

Type of Validity & Examples/Non-Examples

- **Content** The extent to which the content of the test matches the instructional objectives. A semester exam that only includes content covered during the last six weeks is not a valid measure of the course's overall objectives -- it has very low content validity.
- **Criterion** The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) an external criterion. If the end-of-year math tests in 4th grade correlate highly with the state-wide math tests, they would have high concurrent validity.
- **Construct** The extent to which an assessment corresponds to other variables, as predicted by some rationale or theory. If you can correctly hypothesize that Physics students will perform differently on a mechanics test than English students (because of theory), the assessment may have construct validity.

Summary

The real difference between *reliability* and *validity* is mostly a matter of definition.

- Reliability estimates the consistency of your measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects.
- Validity, on the other hand, involves the degree to which your are measuring what you are supposed to, more simply, the accuracy of your measurement. It is my belief that validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).

Threats To Internal Validity

- See main text

Classifying test items

- A method for classifying a test subject in one of a plurality of states in a domain, a domain being a set of facts, a quality measure, or a combination of the two.
- The set of facts for a knowledge domain is any set of facts while the set of facts for a functionality domain is a set of facts relating to the functionality of a test subject.
- A state is characterized by a subset of facts, a value for a quality measure, or a combination of a subset of facts and a value for a quality measure.

Classifying test items

- A first state is higher than or equal to a second state and a second state is lower than or equal to a first state if (1) the subset of facts or the quality measure value associated with the first state respectively includes the subset of facts or is greater than or equal to the quality measure value associated with the second state or (2) the subset of facts and the quality measure value associated with the first state respectively includes the subset of facts and is greater than or equal to the quality measure value associated with the second state.
- Decision-theoretic rules are specified for selecting the test items to be administered to a test subject, for determining when it is appropriate to stop administering test items, and for determining the classification of the test subject. A test subject is classified in the highest state of which he has the knowledge or functionality.

Planning the test

- Test construction, like any other purposeful activity, needs to be adequately planned and executed. The main purpose in the planning stage is to ensure content validity. At this stage preliminary steps that will facilitate the writing of useful and relevant items for specific test types are taken.

Planning the test (contd)

- These steps involve determining the purpose of the test and developing test specifications.
- The decision on the test will depend on the purpose of class testing. A test may be for the purpose of placement or for formative, diagnostic or summative purposes.

Developing test specifications

This involves

- Listing instructional objectives of subject matter for which the test is considered.
- Listing the main topics covered or to be covered by the test
- Marrying objectives and list of topics to construct the table of specifications for the test
- Determining the appropriate test item types

Terminology Used in Testing

The following are some common terms used in testing:

- **Test Blueprint.** The **test** blueprint (or **test** specifications) identifies the objectives and skills which are to be tested and the relative weight on the **test** given to each. This statement necessarily precedes any development of the **test**. These specifications provide a "blueprint" for **test construction**. In absence of such a blueprint, **test** development can potentially proceed with little clear direction. The development of such a set of specifications is the crucial first step in the **test** development process.
- One must be mindful that the **test** specifications cannot and should not remain static. Pedagogy is not static and the specifications for each **test** need to be continually reviewed and modified to reflect the current state of knowledge.

Terminology Used in Testing (contd)

- **Item Development.** The term item is used as a shorthand for questions on the **test**.
- Item development can proceed only when a clearly agreed upon set of objectives is available. To as large an extent as possible, an item should measure only a single objective. Each objective, however, should be measured by one or several items, depending on the **test** specifications.

Terminology Used in Testing (contd)

- **Item format.**

  The format of the item necessarily proceeds from the **test** blueprint. The blueprint indicates the kinds of skills and the balance of **test** content to be measured. The selection of item types and **test** format should be based on the kinds of skills to be measured and not on some personal like or dislike for a particular item format.

Terminology Used in Testing (contd)

- The use of multiple-choice questions, for example, may make sense for large group testing on knowledge of the mechanics of English. This type of item is not generally appropriate, though, as a direct measure of writing skill. If the intent is to determine whether an examinee can write a clear coherent essay, then an essay or free-response format is clearly more appropriate than a multiple-choice format. There is no inherent goodness or badness in any type of question format. The choice must be made on the basis of the behaviour to be tested.

Item format (contd)

- One issue which sometimes constrains the selection of test item format is the need for fast, relatively inexpensive scoring. In general, scoring fixed-response items, such as multiple-choice items, can be done faster and less expensively than scoring free-response items such as fill-in-the-blanks, short answer or essay items. This is particularly true when there are a large number of examinees whose examinations need to be scored quickly.
- Many classroom objectives can be measured adequately with items that are amenable to machine scoring. There are also a number of objectives, however, which are more appropriately measured under other types of formats.
- Instructors (teachers) are encouraged to use or select the type or types of item formats which are best suited for measuring the desired skills.

Terminology Regarding Multiple-Choice Test Questions

- **Multiple-Choice Item:** This is the most common objective-type item. The multiple-choice item is a **test** question which has a number of alternative choices from which the examinee is to select the correct answer. It is generally recommended that one use 4 or 5 choices per question, whenever possible. Using fewer alternatives often results in items with inferior characteristics. The item choices are typically identified on the **test** copy by the letters A through D or E.
- **Stem:** This is the part of the item in which the problem is stated for the examinee. It can be a question, a set of directions or a statement with an embedded blank.
- **Options/Alternatives:** These are the choices given for the item.
- **Key:** This is the correct choice for the item.
- **Distracters:** These are the incorrect choices for the item.

Guidelines for Developing Test Items

The following are some **guidelines** that you should use for preparing **test** items.

Writing Multiple-Choice Test Items

The general rules used for writing multiple-choice items are described below. Recognize that these are general rules; not all rules will be applicable to all types of testing.

1. The stem should contain the problem and any qualifications. The entire stem must always precede the alternatives.
2. Each item should be as short and verbally uncomplicated as possible. Give as much context as is necessary to answer the question, but do not include superfluous information. Be careful not to make understanding the purpose of the item a **test** of reading ability.

Guidelines for Developing Test Items (contd)

3. Avoid negatively stated items. If you have to use this kind of item, emphasize the fact by underlining the negative part, putting it in capital letters or using italics. (For **test construction** purposes, if possible, put all such items together in a single section and indicate this with separate directions.)
4. Keep each item independent from other items. Don't give the answer away to another item. If items require computation, avoid items that are dependent on one another.
5. If one or more alternatives are partially correct, ask for the "best" answer.
6. Try to **test** a different point in each question. If creating item clones (i.e., items designed to measure the exact same aspect of the objective), be certain to sufficiently change the context, vocabulary, and order of alternatives, so that students cannot recognize the two items as clones.

Guidelines for Developing Test Items (contd)

7. If an omission occurs in the stem, it should appear near the end of the stem and not at the beginning.
8. Use a logical sequence for alternatives (e.g., temporal sequence, length of the choice). If two alternatives are very similar (cognitively or visually), they should be placed next to one another to allow students to compare them more easily.
9. Make all incorrect alternatives (i.e., distracters) plausible and attractive. It is often useful to use popular misconceptions and frequent mistakes as distracters. In the foreign languages, item distracters should include only correct forms and vocabulary that actually exists in the language.

Guidelines for Developing Test Items (contd)

10. All alternatives should be homogeneous in content, form and grammatical structure.
11. Use only correct grammar in the stem and alternatives.
12. Make all alternatives grammatically consistent with the stem.
13. The length, explicitness and technical information in each alternatives should be parallel so as not to give away the correct answer.
14. Use **4** or **5** alternatives in each item.

Guidelines for Developing Test Items (contd)

15. Avoid repeating words between the stem and key. It can be done, however, to make distracters more attractive.
16. Avoid wording directly from a reading passage or use of stereotyped phrasing in the key.
17. Alternatives should not overlap in meaning or be synonymous with one another.
18. Avoid terms such as "always" or "never," as they generally signal incorrect choices.

Guidelines for Developing Test Items (contd)

19. To **test** understanding of a term or concept, present the term in the stem followed by definitions or descriptions in the alternatives.
20. Avoid items based on personal opinions unless the opinion is qualified by evidence or a reference to the source of the opinion (e.g., According to the author oft his passage, . .).
21. Do not use "none of the above" as a last option when the correct answer is simply the best answer among the choices offered.
22. Try to avoid "all of the above" as a last option. If an examinee can eliminate any of the other choices, this choice can be automatically eliminated as well.

Writing Essay Test Items

- *Essay items* are useful when examinees have to show how they arrived at an answer.
- A **test** of writing ability is a good example of the kind of **test** that should be given in an essay response format. This type of item, however, is difficult to score reliably and can require a significant amount of time to be graded. Grading is often affected by the verbal fluency in the answer, handwriting, presence or lack of spelling errors, grammar used and the subjective judgements of the grader. Training of graders can require a substantial amount of time and needs to be repeated at frequent intervals throughout the grading.

Writing Essay Test Items (contd)

The following rules may be useful in developing and grading essay questions:

- 1. The shorter the answer required for a given essay item, generally the better. More objectives can be tested in the same period of time, and factors such as verbal fluency, spelling, etc., have less of an opportunity to influence the grader. Help the examinees focus their answers by giving them a starting sentence for their essay.
- 2. Make sure questions are sharply focused on a single issue. Do not give either the examinee or the grader too much freedom in determining what the answer should be.

Guidelines for Writing All Types of Items

- Some additional **guidelines** to consider when writing items are described below:

1. Avoid humorous items. Classroom testing is very important and humorous items may cause students to either not take the exam seriously or become confused or anxious.
2. Items should measure only the construct of interest, not one's knowledge of the item context.
3. Write items to measure what students know, not what they do not know.

Guidelines for Review of Test Items

- The following **guidelines** are recommended for reviewing individual **test** items. When you review an item, write your comments on a copy of the item indicating your suggested changes. If you believe an item is not worth retaining, suggest it be deleted.

1. Consider the **item as a whole** and whether

- a. it measures knowledge or a skill component which is worthwhile and appropriate for the examinees who will be tested;
- b. there is a markedly better way to **test** what this item tests;
- c. it is of the appropriate level of difficulty for the examinees who will be tested.

Guidelines for Review of Test Items (contd)

2. Consider the **stem** and whether it

- a. presents a clearly defined problem or task to the examinee;
- b. contains unnecessary information;
- c. could be worded more simply, clearly or concisely.

3. Consider the **alternatives** and whether

- a. they are parallel in structure;
- b. they fit logically and grammatically with the stem;
- c. they could be worded more simply, clearly or concisely;
- d. any are so inclusive that they logically eliminate another more restricted option from being a possible answer.

Guidelines for Review of Test Items (contd)

4. Consider the **key** and whether it

- a. is the best answer among the set of options for the item;
- b. actually answers the question posed in the stem;
- c. is too obvious relative to the other alternatives (i.e., should be shortened, lengthened, given greater numbers of details, made less concrete).

5. Consider the **distracters** and whether

- a. there is any way you could justify one or more as an acceptable correct answer;
- b. they are plausible enough to be attractive to examinees who are misinformed or ill-prepared;
- c. any one calls attention to the key (e.g., no distracter should merely state the reverse of the key or resemble the key very closely unless another pair of choices is similarly parallel or involves opposites).

Assembling a Test Form

General Rules for Test Assembly

- The following are general rules, intended as **guidelines** for assembling test forms. When reviewing a **test** prior to administering, verify that the **test** conforms with the following **test construction guidelines**.

Test Construction Rules for Multiple-Choice Tests.

1. Set the number of items so that at least 95 percent of the examinees can answer all items.
2. The correct choice should appear about an equal number of times in each response position.
3. Do not use any pattern of correct responses, e.g., ABCDE, etc.
4. Directions to examinees should be written on the **test** to indicate whether guessing is permitted or not.

Test Construction Rules for Essay Tests.

- All examinees must take the same items. Do not give them a chance to choose which items they want to answer. Meaningful comparisons normally can be made only if all examinees take the same **test**.

Grading Essay Tests

- Because of their subjective nature, essay exams are difficult to grade. The following **guidelines** are helpful for grading essay exams in a consistent and meaningful way.

1. Construct a model answer for each item and award a point for each essential element of the model answer. This should help minimize the subjective effects of grading.
2. Essay items must be graded anonymously if at all possible in order to reduce the subjectivity of the graders. That is, graders should not be informed as to the identity of the examinees whose papers they are grading.

Grading Essay Tests (contd)

3. Grade a single essay item at a time. This helps the grader maintain a single set of criteria for awarding points to the response. In addition, it tends to reduce the influence of the examinee's previous performance on other items.
4. Unless it is a **test** of language mechanics, do not take off credit for poor handwriting, spelling errors, poor grammar, failure to punctuate properly, etc.
5. Ideally, there should be two graders for each item. Any disagreements between these two graders must be resolved by a third grader. Normally, this third grader is the head grader or course instructor.

Guidelines for Test Design and Construction

- Step 1: Defining the constructs you want to measure and outline the proposed content of the Test

- **Aptitude tests for job applicants** :

- First conduct a ***job analysis*** (task analysis) : listing the important components of the position you are trying to fill.

Guidelines for Test Design and Construction (contd)

- The **Job analysis** will contain the **critical incidents,**
- A list of work related behaviors which are essential for successful completion of the job.

- A well designed aptitude test will contain items which measure the entire cross-section of critical incidents.

- To paraphrase, a properly constructed test will measure a ***representative sample*** of most critical incidents.

Test Planners must consider a variety of issues

1. What are the topics and materials to be tested ?
2. What Kind of Questions should be constructed?
3. What item and test formats should be used ?
4. When, where, and how is the test to be given ?
5. How should the tests be scored ?

Essay Tests vs. Multiple Choice Tests

Advantages of Essay Tests

- Trivially Easy to construct an essay test, compared to the time required to construct a multiple choice test.
- Essay Tests allow for the examination of higher order cognitive objectives.
- Allows test takers to show the depth of knowledge they have of a particular subject area.
- Allows test takers to practice their writing skills.

Disadvantages of Essay Tests

- Subjectivity in grading means it is possible for two people evaluating the same essay to calculate different grades.
- Grading essays takes a significant amount of time (may not be feasible for tests of a large group of people).
- Because of time it takes to write essay, may not be able to survey a large portion of the subject material. (problem with representative sampling)
- Students may answer a different question than asked.

Advantages of Multiple Choice Tests

- Objective scoring procedures means anyone can score a particular exam and come up with the same grade. (helps to increase reliability of the test)
- A representative sample of questions from all subject areas to be tested can be easily obtained.
- Exams can be scored relatively quickly, very suitable for large numbers of test takers.

Disadvantages of Multiple Choice Tests

- Take considerable time to create compared to essay tests.
- Much more difficult to assess higher order cognitive objectives such as *analysis* and *evaluation*.
- Can be made unintentionally more difficult by the use of : **Negatives** or (even worse) double negatives within the question and **Interlocked Items** : Where you must get the correct answer to a preceding question in order to get the right answer to the current question. Can be made easier by the use of **interrelated items**.

Standard Multiple Choice Formats

Short Answer Questions : Test taker must supply the answer.

- **<u>3 Major Guidelines :</u>**
- **Questions are preferable to incomplete statements**
- **If fill-in-the-blank format is used, the blank should come at the end of the statement.**
- **Avoid multiple blanks in the same item.**
- The _____ is a good example of a _____ test.

True-False Items

Some of the most commonly used and simplest to create test items. True false items are sometimes criticized for encouraging rote memorization.

- True-False answers are affected by the use of *specific determiners* such as **never, always, and only,** indicating a statement which is usually false.

  Often, sometimes, and usually commonly indicate a true statement

- **Guidelines for constructing good True-False statements :**

1. Ensure the statements deal with non-trivial information to discourage rote memorization.
2. Keep statements short in length and unambiguous.
3. Avoid negatively stated items, especially double negatives.
4. Avoid specific Determiners such as always, never, and only.

True-False Items (contd)

5. On Opinion statements, cite the source of the opinion.
6. Avoid tricky items.
7. Make true and false statements about equal length, and include an equal amount of both.
8. Make wrong answers more attractive by wording items in such a way that they are not obviously wrong.

Guidelines for Matching Items

- Matching items are reasonably easy to construct.
- The drawback is that matching items primarily test lower order cognitive objectives, and thus encourage rote memorization.

5 Guidelines for Matching Items

1. Arrange the premise and response options in a clear logical column format, with question stem on left hand side and to be matched items on the right.
2. Number the question stems sequentially, and use letters to differentiate among response choices.
3. Use between 6 to 15 question stems, and two to three more response options than question stems.

Guidelines for Matching Items (contd)

4. Clearly specify whether the matching is one-to-one, one-to-many, and whether or not response choices can be used more than once.
5. Put the entire matching section of a single test page for clarity.

Ranking and *rearrangement* items are special type of matching questions in which order counts.

Multiple – Choice Questions

- Are very versatile as you can measure the attainment of both lower and higher order cognitive objectives.
- Scores on Multiple choice questions are less affected by response bias (response set) than are true-false items.
- The key to constructing a high quality multiple choice question is to have good, plausible distracters.

Shortcomings of multiple choice items

1. Good items take time and thought to construct
2. Multiple choice stresses recognition over recall. -Familiarity with an answer can lead to correct guessing.
3. Require more time to answer than true-false questions.

Guidelines for constructing multiple choice items

1. Question format preferred to incomplete statement. If incomplete statement is used, blank should be at end of statement.
2. State question clearly and at the appropriate reading level.
3. Place as much of the item is the stem as possible. Answer choices should be as short as possible.

Guidelines for constructing multiple choice items (contd)

4. Use opinion questions sparingly and cite the source or authority of the opinion.
5. **Four** to five choices are standard, but two or three choices can be used as well.

   How many choices appear is partially dependent on the ease of generating high quality distracters

Guidelines for constructing multiple choice items (contd)

6. If the answer choices have a natural order, arrange them in that fashion. (Dates, ages). Otherwise answer choices should be randomly arranged.
7. Try to make answer choices equal in length and complexity.

Guidelines for constructing multiple choice items (contd)

8. Try to make all answer choices plausible, but only one correct answer.
9. Formulate a logical reason why someone who doesn't know the correct answer would select from the distracter set.
10. Avoid the use of negatives, and in particular, double negatives.

Guidelines for constructing multiple choice items (contd)

11. Ambiguous and tricky options should be avoided.
12. Specific determiners should be avoided, and “all of the above” and ‘none of the above’ questions should be infrequent.
13. Place options in Stacked format, rather than row by row.

Guidelines for constructing multiple choice items (contd)

14. Make sure the amount of items tested is appropriate for the time constraints of the testing session.
15. Item difficulty should be such that overall performance is halfway between chance (pure guessing) and 100%. This will give your testing measure the maximum ability to separate according to performance.

Constructing Distracter Items

- Usually, the question and the answer that you want to ask is relatively easy to develop.
- But what often takes the most time is coming up with plausible distracters.

Two approaches can be taken :

- The Rational Approach : The test developers' understanding of the subject material and their ability to organize that material leads them to adopt specific distracters for specific test items

Constructing Distracter Items (contd)

- The Empirical Approach : **You select distracters based on pre-test data.**

  You administer the test without any answer choices and have people complete the measure as a short answer test.
- Your test subjects should be similar in composition to the population you will actually be testing.
- You compile a list of incorrect answers for each question, and use the most popular (most frequent) incorrect answer as your distracter choices when you reassemble the exam.

What is item analysis?

- is a process of examining class-wide (classroom) performance on individual test items. There are three common types of item analysis which provide teachers with three different types of information:

Difficulty Index

- -Teachers produce a difficulty index for a test item by calculating the *proportion of students in class who got an item correct*. (The name of this index is counter-intuitive, as one actually gets a measure of how easy the item is, not the difficulty of the item.) The larger the proportion, the more students who have learned the content measured by the item.

Discrimination Index

- The discrimination index is a basic measure of the validity of an item.
- It is a measure of an item's ability to discriminate between those who scored high on the total test and those who scored low.

Analysis of Response Options

- In addition to examining the performance of an entire test item, teachers are often interested in examining the performance of individual distracters (incorrect answer options) on multiple-choice items.

Analysis of Response Options (contd)

- By calculating the proportion of students who chose each answer option, teachers can identify which distracters are "working" and appear attractive to students who do not know the correct answer, and which distracters are simply taking up space and not being chosen by many students.

Analysis of Response Options (contd)

- To eliminate blind guessing which results in a correct answer purely by chance (which hurts the validity of a test item), teachers want as many plausible distracters as is feasible. Analyses of response options allow teachers to fine tune and improve items they may wish to use again with future classes.

Discrimination Index (contd)

- Though there are several steps in its calculation, once computed, this index can be interpreted as an indication of the extent to which overall knowledge of the content area or mastery of the skills is related to the response on an item. Perhaps the most crucial validity standard for a test item is that whether a student got an item correct or not is due to their level of knowledge or ability and not due to something else such as chance or test bias.

Analysis of Response Options (contd)

- To eliminate blind guessing which results in a correct answer purely by chance (which hurts the validity of a test item), teachers want as many plausible distracters as is feasible. Analyses of response options allow teachers to fine tune and improve items they may wish to use again with future classes.

Example

- See text

Interpreting test scores

- Mean
- Median
- Mode
- Variance
- Standard devian
- Grouped data